A. Derivation of regularized ERM duality

For completeness, in this section we derive the dual (5) to the problem of computing proximal operator for the ERM objective (3).

We can rewrite the primal problem as

$$\begin{array}{ll}\min\limits_{x\in\mathbb{R}^d, z\in\mathbb{R}^n} & \sum_{i=1}^n \phi_i(z_i) + \frac{\lambda}{2}\|x-s\|_2^2 \\ \text{subject to} & z_i = a_i^\mathsf{T} x, \quad \text{for } i = 1,\dots,n\end{array}.$$

By convex duality, this is equivalent to

$$\begin{aligned}\min_{x,\{z_i\}} \max_{y\in\mathbb{R}^n} \sum_{i=1}^n \phi_i(z_i) + \frac{\lambda}{2}\|x-s\|_2^2 + y^\mathsf{T}(Ax-z) &= \max_y \min_{x,\{z_i\}} \sum_{i=1}^n \phi_i(z_i) + \frac{\lambda}{2}\|x-s\|_2^2 + y^\mathsf{T}(Ax-z) \\ &= \max_y \min_{x,\{z_i\}} \sum_{i=1}^n (\phi_i(z_i) - y_i z_i) + \frac{\lambda}{2}\|x-s\|_2^2 + y^\mathsf{T} Ax \\ &= \max_y \sum_{i=1}^n \min_{z_i} \{\phi_i(z_i) - y_i z_i\} + \min_x \left\{\frac{\lambda}{2}\|x-s\|_2^2 + y^\mathsf{T} Ax\right\} \\ &= \max_y \sum_{i=1}^n -\max_{z_i} \{y_i z_i - \phi_i(z_i)\} - \max_x \left\{-y^\mathsf{T} Ax - \frac{\lambda}{2}\|x-s\|_2^2\right\} \\ &= \max_y \sum_{i=1}^n -\phi_i^*(y_i) - \frac{1}{2\lambda}\|A^\mathsf{T} y\|_2^2 + s^\mathsf{T} A^\mathsf{T} y \\ &= -\min_y \sum_{i=1}^n \phi_i^*(y_i) + \frac{1}{2\lambda}\|A^\mathsf{T} y\|_2^2 - s^\mathsf{T} A^\mathsf{T} y.\end{aligned}$$

The final negated problem is precisely the dual formulation.

The first problem is a Lagrangian saddle-point problem, where the Lagrangian is defined as

$$\mathcal{L}(x,y,z) = \sum_{i=1}^n \phi_i(z_i) + \frac{\lambda}{2}\|x-s\|_2^2 + y^\mathsf{T}(Ax-z).$$

The dual-to-primal mapping (6) and primal-to-dual mapping (7) are implied by the KKT conditions under $\mathcal{L}$, and can be derived by solving for x, y, and z in the system $\nabla\mathcal{L}(x,y,z) = 0$.

The *duality gap* in this context is defined as

$$\text{gap}_{s,\lambda}(x,y) = f_{s,\lambda}(x) + g_{s,\lambda}(y). \tag{9}$$

Strong duality dictates that $\text{gap}_{s,\lambda}(x,y) \geq 0$ for all $x \in \mathbb{R}^d$, $y \in \mathbb{R}^n$, with equality attained when x is primal-optimal and y is dual-optimal.

B. Additional lemmas

This section establishes technical lemmas, useful in statements and proofs throughout the paper.

Lemma B.1 (Standard bounds for smooth, strongly convex functions)**.** *Let $f : \mathbb{R}^k \to \mathbb{R}$ be differentiable and let $x \in \mathbb{R}^k$. Furthermore, let x^{opt} be a minimizer of f.*

If f is L-smooth then

$$\frac{1}{2L}\|\nabla f(x)\|_2^2 \leq f(x) - f(x^{opt}) \leq \frac{L}{2}\|x - x^{opt}\|_2^2$$

If f is μ-strongly convex we have

$$\frac{\mu}{2}\|x - x^{opt}\|_2^2 \leq f(x) - f(x^{opt}) \leq \frac{1}{2\mu}\|\nabla f(x)\|_2^2$$

Proof. Apply the definition of smoothness and strong convexity at the points x and x^{opt} and minimize the resulting quadratic form. □

Lemma B.2 (Errors for primal-dual pairs)**.** *Consider F, $f_{s,\lambda}$ $g_{s,\lambda}$, $\hat{x}_{s,\lambda}$, and $\hat{y}$, as defined in (1), (3), (5), (6), and (7), respectively. Then for all $y \in \mathbb{R}^n$,*

$$f_{s,\lambda}(\hat{x}_{s,\lambda}(y)) - f_{s,\lambda}^{opt} \leq \frac{R^2L(nR^2L + \lambda)}{\lambda^2}(g_{s,\lambda}(y) - g_{s,\lambda}^{opt}),$$

Furthermore let

$$x_{s,\lambda}^{opt} = \operatorname{argmin} f_{s,\lambda}(x) \quad \textit{and} \quad y_{s,\lambda}^{opt} = \operatorname{argmin} g_{s,\lambda}(x).$$

Then for all $y \in \mathbb{R}^n$,

$$\|\hat{x}_{s,\lambda}(y) - x_{s,\lambda}^{opt}\|_2^2 \leq \frac{nR^2}{\lambda^2}\|y - y_{s,\lambda}^{opt}\|_2^2.$$

Proof. Because F is nR^2L smooth, $f_{s,\lambda}$ is $nR^2L + \lambda$ smooth. Consequently, for all $x \in \mathbb{R}^d$ we have

$$f_{s,\lambda}(x) - f_{s,\lambda}^{\text{opt}}(x_{s,\lambda}^{\text{opt}}) \leq \frac{nR^2L + \lambda}{2}\|x - x_{s,\lambda}^{\text{opt}}\|_2^2$$

Since we know that $x_{s,\lambda}^{\text{opt}} = s - \frac{1}{\lambda}A^\mathsf{T}y_{s,\lambda}^{\text{opt}}$ and $AA^\mathsf{T} \preceq nR^2I$,

$$\begin{aligned} f_{s,\lambda}(\hat{x}_{x,\lambda}(y)) - f_{s,\lambda}(x_{s,\lambda}^{\text{opt}}) &\leq \frac{nR^2L + \lambda}{2}\|s - \frac{1}{\lambda}A^\mathsf{T}y - (s - \frac{1}{\lambda}A^\mathsf{T}y_{s,\lambda}^{\text{opt}})\|_2^2 \\ &= \frac{nR^2L + \lambda}{2\lambda^2}\|y - y_{s,\lambda}^{\text{opt}}\|_{AA^\mathsf{T}}^2 \\ &\leq \frac{nR^2(nR^2L + \lambda)}{2\lambda^2}\|y - y_{s,\lambda}^{\text{opt}}\|_2^2. \end{aligned} \tag{10}$$

Finally, since each ϕ_i^* is $1/L$ strongly convex, G is n/L strongly convex and hence so is $g_{s,\lambda}$. Therefore,

$$\frac{n}{2L}\|y - y_{s,\lambda}^{\text{opt}}\|_2^2 \leq g_{s,\lambda}(y) - g_{s,\lambda}(y_{s,\lambda}^{\text{opt}}). \tag{11}$$

Substituting (11) in (10) yields the result. □

By adding the dual error to both sides of the inequality in Lemma B.2, we obtain the following corollary.

Corollary B.3 (Dual error bounds gap)**.** *Consider $g_{s,\lambda}$, $\text{gap}_{s,\lambda}$, $\hat{x}_{s,\lambda}$, and $\hat{y}$, as defined in (5), (9), (6), and (7), respectively. For all $s \in \mathbb{R}^d$ and $y \in \mathbb{R}^n$,*

$$\text{gap}_{s,\lambda}(\hat{x}_{s,\lambda}(y), y) \leq \frac{2R^2L(nR^2L + \lambda)}{\lambda^2}(g_{s,\lambda}(y) - g_{s,\lambda}^{opt}).$$

Another corollary arises by combining Lemmas B.1 and B.3.

Corollary B.4 (Dual error bounds primal gradient)**.** *Consider $f_{s,\lambda}$, $g_{s,\lambda}$, and $\hat{x}_{s,\lambda}$, as defined in (3), (5), and (6), respectively. For all $s \in \mathbb{R}^d$ and $y \in \mathbb{R}^n$,*

$$\begin{aligned} \|\nabla f_{s,\lambda}(\hat{x}_{s,\lambda}(y))\|_2^2 &\leq 2(nR^2L + \lambda)(f_{s,\lambda}(\hat{x}_{s,\lambda}(y)) - f_{s,\lambda}^{opt}) \\ &\leq \frac{2R^2L(nR^2L + \lambda)^2}{\lambda^2}(g_{s,\lambda}(y) - g_{s,\lambda}^{opt}). \end{aligned}$$

Lemma B.5 (Gap for primal-dual pairs)**.** *Consider F, $f_{s,\lambda}$ $g_{s,\lambda}$, $\mathrm{gap}_{s,\lambda}$, $\hat{x}_{s,\lambda}$, and $\hat{y}$, as defined in (1), (3), (5), (9), (6), and (7), respectively. For any $x, s \in \mathbb{R}^d$*

$$\mathrm{gap}_{s,\lambda}(x, \hat{y}(x)) = \frac{1}{2\lambda}\|\nabla F(x)\|_2^2 + \frac{\lambda}{2}\|x - s\|_2^2. \tag{12}$$

Separately, for any $y \in \mathbb{R}^n$ and $s \in \mathbb{R}^d$,

$$\mathrm{gap}_{s,\lambda}(\hat{x}_{s,\lambda}(y), y) = F(\hat{x}_{s,\lambda}(y)) + g_{s,\lambda}(y) + \tfrac{1}{2\lambda}\|A^\mathsf{T} y\|^2. \tag{13}$$

Proof. To prove the first identity (12), let $\hat{y} = \hat{y}(x)$ for brevity. Recall that

$$\hat{y}_i = \phi_i'(a_i^\mathsf{T} x) \in \operatorname*{argmax}_{y_i}\{x^\mathsf{T} a_i y_i - \phi_i^*(y_i)\} \tag{14}$$

by definition, and hence $x^\mathsf{T} a_i \hat{y}_i - \phi_i^*(\hat{y}_i) = \phi_i(a_i^\mathsf{T} x)$. Observe that

$$\begin{aligned}
\mathrm{gap}_{s,\lambda}(x, \hat{y}) &= \sum_{i=1}^n \left(\phi_i(a_i^\mathsf{T} x) + \phi_i^*(\hat{y}_i)\right) - x^\mathsf{T} A^\mathsf{T} \hat{y} + \tfrac{1}{2\lambda}\|A^\mathsf{T}\hat{y}\|^2 + \tfrac{\lambda}{2}\|x - s\|^2 \\
&= \sum_{i=1}^n \left(\underbrace{\phi_i(a_i^\mathsf{T} x) + \phi_i^*(\hat{y}_i) - x^\mathsf{T} a_i \hat{y}_i}_{=0 \text{ (by (14))}} \right) + \tfrac{1}{2\lambda}\|A^\mathsf{T}\hat{y}\|^2 + \tfrac{\lambda}{2}\|x - s\|^2 \\
&= \tfrac{1}{2\lambda}\|A^\mathsf{T}\hat{y}\|^2 + \tfrac{\lambda}{2}\|x - s\|^2 \\
&= \tfrac{1}{2\lambda}\|\sum_{i=1}^n a_i \phi_i'(a_i^\mathsf{T} x)\|^2 + \tfrac{\lambda}{2}\|x - s\|^2 \\
&= \tfrac{1}{2\lambda}\|\nabla F(x)\|^2 + \tfrac{\lambda}{2}\|x - s\|^2.
\end{aligned}$$

For the second identity (13), let $\hat{x} = \hat{x}_{s,\lambda}(y)$ for brevity. Fix $s \in \mathbb{R}^d$ and $\lambda > 0$. Define $r(x) = \frac{\lambda}{2}\|x - s\|^2$, and note that its Fenchel conjugate is $r^*(y) = \frac{1}{2\lambda}\|y\|^2 + s^\mathsf{T} y$. With this notation we can write:

$$\begin{aligned}
f_{s,\lambda}(x) &= F(x) + r(x) \\
g_{s,\lambda}(y) &= G(y) + r^*(-A^\mathsf{T} y).
\end{aligned}$$

Observe that

$$\begin{aligned}
\hat{x} &= s - \tfrac{1}{\lambda} A^\mathsf{T} y \\
&= \operatorname*{argmin}_x \left\{ \tfrac{\lambda}{2}\|x - s\|^2 + x^\mathsf{T} A^\mathsf{T} y \right\} \\
&= \operatorname*{argmin}_x \left\{ r(x) + x^\mathsf{T} A^\mathsf{T} y \right\} \\
&= \operatorname*{argmax}_x \left\{ -x^\mathsf{T} A^\mathsf{T} y - r(x) \right\} \\
&= \nabla r^*(-A^\mathsf{T} y).
\end{aligned}$$

Note also that $g_{s,\lambda}$ may be rewritten as

$$\begin{aligned}
g_{s,\lambda}(y) &= G(y) + \tfrac{1}{2\lambda}\|A^\mathsf{T} y\|^2 - s^\mathsf{T} A^\mathsf{T} y \\
&= G(y) + (-\tfrac{1}{2\lambda} y^\mathsf{T} A + \tfrac{1}{\lambda} y^\mathsf{T} A - s^\mathsf{T}) A^\mathsf{T} y \\
&= G(y) - \tfrac{1}{2\lambda}\|A^\mathsf{T} y\|^2 - \hat{x}^\mathsf{T} A^\mathsf{T} y.
\end{aligned}$$

Combining the above two observations, and noting that the first implies equality in the Fenchel-Young inequality,

$$\begin{aligned}
\mathrm{gap}_{s,\lambda}(\hat{x}, y) &= f_{s,\lambda}(\hat{x}) + g_{s,\lambda}(y) \\
&= F(\hat{x}) + G(y) + r(\hat{x}) + r^*(-A^\mathsf{T} y) \\
&= F(\hat{x}) + G(y) - \hat{x}^\mathsf{T} A^\mathsf{T} y \\
&= F(\hat{x}) + g_{s,\lambda}(y) + \tfrac{1}{2\lambda}\|A^\mathsf{T} y\|^2,
\end{aligned}$$

proving the claim. □

C. Convergence analysis of Dual APPA

The goal of this section is to establish the convergence rate of Algorithm 3, Dual APPA. It is structured as follows:

- Mirroring Lemma 3.2, we establish Lemma C.1, concerning contraction of the norm of the gradient, $\|\nabla F\|$, rather than of the error in function value $F - F^{\text{opt}}$. Similar to Lemma 3.2, this lemma is self-contained and assume only that F is μ-strongly convex, but not that it is the ERM objective.
- We then establish directly a geometric rate upper bound on the value of $\|\nabla F\|$ over the course of the algorithm, relying, in the process on Lemma C.1.
- Finally, we query the strong convexity of F to establish that it too is subject to the same rate (up to an extra multiplicative factor).

Lemma C.1 (Gradient-norm reduction implies contraction)**.** *Suppose $F : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and that $\lambda > 0$. Define $f_s : \mathbb{R}^d \to \mathbb{R}$ by $f_s(x) = f(x) + \frac{\lambda}{2}\|x - s\|_2^2$. For any $x_0, x_1 \in \mathbb{R}^d$,*

$$\|\nabla F(x_1)\|_2 \leq \left(1 + \frac{\lambda}{\lambda + \mu}\right) \|\nabla f_{x_0}(x_1)\|_2 + \frac{\lambda}{\lambda + \mu}\|\nabla F(x_0)\|_2. \tag{15}$$

Corollary C.2. *Define $\gamma \stackrel{\text{def}}{=} \lambda/(\lambda + \mu)$ and let τ be any scalar in the interval $[\gamma, 1)$. For any $x_0, x_1 \in \mathbb{R}^d$,*

$$\|\nabla F(x_1)\| \leq \tau \|\nabla F(x_0)\|, \tag{16}$$

whenever

$$\|\nabla f_{x_0}(x_1)\| \leq \left(\frac{\tau - \gamma}{1 + \gamma}\right) \|\nabla F(x_0)\|. \tag{17}$$

Proof of Lemma C.1. Taking gradients of f_{x_0} and f_{x_1} we have that

$$\begin{aligned} \nabla F(x_1) &= \nabla F(x_1) + \lambda(x_1 - x_0) - \lambda(x_1 - x_0) \\ &= \nabla f_{x_0}(x_1) - \lambda(x_1 - x_0). \end{aligned}$$

By the triangle inequality,

$$\|\nabla F(x_1)\| \leq \|\nabla f_{x_0}(x_1)\| + \lambda\|x_1 - x_0\|.$$

Because f_{x_0} is $(\lambda + \mu)$-strongly convex,

$$\begin{aligned} (\lambda + \mu)\|x_1 - x_0\| &\leq \|\nabla f_{x_0}(x_1) - \nabla f_{x_0}(x_0)\| \\ &\leq \|\nabla f_{x_0}(x_1)\| + \|\nabla f_{x_0}(x_0)\| \\ &= \|\nabla f_{x_0}(x_1)\| + \|\nabla F(x_0)\|. \end{aligned}$$

Combining the previous two observations proves the claim. □

Throughout the remainder of this section, we consider the functions defined in the main setup (Section 1.1) – namely, F, $f_{s,\lambda}$ $g_{s,\lambda}$, $\hat{x}_{s,\lambda}$, and $\hat{y}$, as defined in (1), (3), (5), (6), and (7), respectively, where F is a μ-strongly convex sum whose constituent summands are each L-smooth scalar functions operating on the inner product of the variable $x \in \mathbb{R}^d$ with a vector a_i of Euclidean norm at most R. For notational brevity, we omit the λ subscript, and we denote iterates of the algorithm with numerical subscripts (*e.g.* $x_1, x_2, \ldots$, and $y_1, y_2, \ldots$) rather than parenthesized superscripts.

We begin by bounding the error of the dual regularized ERM problem when the center of regularization changes. This characterizes the initial error at the beginning of each Dual APPA iteration.

Lemma C.3 (Dual error is bounded across re-centering.)**.** *Fix $y \in \mathbb{R}^n$ and $x \in \mathbb{R}^d$. Let $x' = \hat{x}_x(y)$. Then*

$$g_{x'}(y) - g_{x'}^{opt} \leq c_1(g_x(y) - g_x^{opt}) + c_2\|\nabla F(x)\|_2^2,$$

where

$$c_1 = \frac{2R^2L(nR^2L+\lambda)}{\lambda^2}\left(1 + \frac{\lambda(nR^2L+\lambda)}{2(\lambda+\mu)^2}\right) \quad \textit{and} \quad c_2 = \frac{\lambda}{2(\lambda+\mu)^2}.$$

In other words, the dual error $g_s(y) - g_s^{\text{opt}}$ is bounded across a re-centering step by multiples of previous sub-optimality measurements (namely, dual error and gradient norm).

Proof. We first show how re-centering, from x to x', increases the duality gap between y and x' (the new center). The dual function value changes from

$$g_x(y) = G(y) + \frac{1}{2\lambda}\|A^\mathsf{T} y\|^2 - x^\mathsf{T} A^\mathsf{T} y$$

to

$$\begin{aligned}
g_{x'}(y) &= G(y) + \frac{1}{2\lambda}\|A^\mathsf{T} y\|^2 - x'^\mathsf{T} A^\mathsf{T} y \\
&= G(y) + \frac{1}{2\lambda}\|A^\mathsf{T} y\|^2 - \left(x - \frac{1}{\lambda}A^\mathsf{T} y\right)^\mathsf{T} A^\mathsf{T} y \\
&= g_x(y) + \frac{1}{\lambda}\|A^\mathsf{T} y\|^2 \\
&= g_x(y) + \lambda\|\frac{1}{\lambda}A^\mathsf{T} y\|^2 \\
&= g_x(y) + \lambda\|x - \frac{1}{\lambda}A^\mathsf{T} y - x\|^2 \\
&= g_x(y) + \lambda\|x' - x\|^2
\end{aligned}$$

i.e. it increases by $\lambda\|x' - x\|^2$. Meanwhile, the primal function value changes from

$$f_x(x') = F(x') + \frac{\lambda}{2}\|x' - x\|^2$$

to $f_{x'}(x') = F(x')$, *i.e.* it decreases by $\frac{\lambda}{2}\|x' - x\|^2$. Hence, in total, the new duality gap is

$$\begin{aligned}
\text{gap}_{x'}(x', y) &= \text{gap}_x(x', y) + \frac{\lambda}{2}\|x' - x\|^2 \\
&\leq \text{gap}_x(x', y) + \frac{\lambda}{2(\mu+\lambda)^2}\|\nabla f_x(x') - \nabla f_x(x)\|^2 \\
&\leq \text{gap}_x(x', y) + \frac{\lambda}{2(\mu+\lambda)^2}\left(\|\nabla f_x(x')\|^2 + \|\nabla f_x(x)\|^2\right) \\
&\leq \text{gap}_x(x', y) + \frac{\lambda}{2(\mu+\lambda)^2}\left(\|\nabla f_x(x')\|^2 + \|\nabla F(x)\|^2\right),
\end{aligned}$$

where the first inequality follows by $(\mu+\lambda)$-strong convexity of f_x.

Combining the last two chains of inequalities, and abbreviating $\tilde{L} = nR^2L + \lambda$, we bound the re-centered dual error of y:

$$\begin{aligned}
g_{x'}(y) - g_{x'}^{\text{opt}} &\leq \text{gap}_{x'}(x', y) \\
&\leq \text{gap}_x(x', y) + \frac{\lambda}{2(\mu+\lambda)^2}\left(\|\nabla f_x(x')\|^2 + \|\nabla F(x)\|^2\right) \\
&\leq \frac{2R^2L\tilde{L}}{\lambda^2}(g_x(y) - g_x^{\text{opt}}) + \frac{\lambda}{2(\mu+\lambda)^2}\left(\|\nabla f_x(x')\|^2 + \|\nabla F(x)\|^2\right) \\
&\leq \frac{2R^2L\tilde{L}}{\lambda^2}(g_x(y) - g_x^{\text{opt}}) + \frac{\lambda}{2(\mu+\lambda)^2}\left(\frac{2R^2L\tilde{L}^2}{\lambda^2}(g_x(y) - g_x^{\text{opt}}) + \|\nabla F(x)\|^2\right) \\
&= \frac{2R^2L\tilde{L}}{\lambda^2}\left(1 + \frac{\lambda\tilde{L}}{2(\mu+\lambda)^2}\right)(g_x(y) - g_x^{\text{opt}}) + \frac{\lambda}{2(\mu+\lambda)^2}\|\nabla F(x)\|^2.
\end{aligned}$$

where the third and fourth inequalities follow from Corollary B.3 – the fourth first relying on the $\tilde{L}$-smoothness of f_x, that is:

$$\|\nabla f_x(x')\|^2 \leq 2\tilde{L}(f_x(x') - f_x^{\text{opt}}) \leq \frac{2R^2L\tilde{L}^2}{\lambda^2}(g_x(y) - g_x^{\text{opt}}). \tag{18}$$

Recalling the choices of numerical constants c_1 and c_2, the claim is proven. □

Finally, we prove that – for an appropriate choice of oracle parameter σ independent of the target error ϵ – the iterates produced by Dual APPA are bounded above by a geometric sequence. In doing so, it will be convenient to abbreviate

$$c_3 = \frac{2R^2L(nR^2L + \lambda)^2}{\lambda^2},$$

the same numerical constant as in (18).

Proposition C.4 (Convergence rate of Dual APPA in gradient norm)**.** *Define $\gamma = \lambda/(\lambda + \mu)$ and let τ be any scalar in $[\gamma, 1)$. Define*

$$r = \max\left\{ 1/2,\ \tau/\sqrt{2},\ \tau^2 \right\}, \tag{19}$$

and

$$\sigma = \max\left\{ \frac{c_3}{2\lambda}\left(\frac{\tau-\gamma}{1+\gamma}\right)^{-2},\ \frac{c_1 + c_2 + c_2c_3}{r},\ \frac{1}{r}\left(\frac{\tau-\gamma}{1+\gamma}\right)^{-2} c_3(c_1 + c_2 + c_2c_3) \right\}. \tag{20}$$

In the execution of Dual APPA (Algorithm 3), for every iteration $t \geq 1$,

$$g_{x_{t-1}}(y_t) - g_{x_{t-1}}^{opt} \leq cr^t, \quad \textit{and} \tag{21}$$

$$\|\nabla F(x_{t-1})\|^2 \leq cr^t, \tag{22}$$

where $c = 2\|\nabla F(x_0)\|^2$.

Proof. The argument proceeds by strong induction on t. We begin the inductive argument with base case $t = 1$.

For the first invariant (21), the algorithm takes $\sigma \geq 1/(2\lambda)$ and we have

$$g_{x_0}(y_1) - g_{x_0}^{\text{opt}} \leq \frac{1}{\sigma}(g_{x_0}(y_0) - g_{x_0}^{\text{opt}}) \leq \frac{1}{\sigma}\frac{1}{2\lambda}\|\nabla F(x_0)\|^2 \leq \frac{1}{\sigma}\frac{1}{4\lambda}c \tag{23}$$

because the algorithm explicitly takes $y_0 = \hat{y}(x_0)$ (see Lemma B.4). So this part of the claim holds as $r \geq 1/2$ and $\sigma \geq 1/(2\lambda)$.

The second invariant (22) holds immediately as $c = 2\|\nabla F(x_0)\|^2$ and $r \geq 1/2$.

We will also explicitly handle the case $t = 2$ for the second invariant. By (23) and our choice of σ, we have that

$$\|\nabla f_{x_0}(x_1)\|^2 \leq c_3(g_{x_0}(y_1) - g_{x_0}^{\text{opt}}) \leq c_3 \frac{1}{\sigma}\frac{1}{2\lambda}\|\nabla F(x_0)\|^2 \leq \left(\frac{\tau-\gamma}{1+\gamma}\right)^2 \|\nabla F(x_0)\|^2.$$

and so by Corollary C.2,

$$\|\nabla F(x_1)\|^2 \leq \tau^2\|\nabla F(x_0)\|^2 \leq \frac{\tau^2}{2}c.$$

and the claim follows as $r^2 \geq \tau^2/2$

Now consider $t \geq 2$. Concerning the first invariant (21), the algorithm takes $\sigma \geq \frac{1}{r}(c_1 + c_2 + c_2c_3)$ and so we have

$$\begin{aligned}
g_{x_{t-1}}(y_t) - g_{x_0}^{\text{opt}} &\leq \frac{1}{\sigma}(g_{x_{t-1}}(y_{t-1}) - g_{x_{t-1}}^{\text{opt}}) \\
&\leq \frac{1}{\sigma}\left[(c_1 + c_2c_3)(g_{x_{t-2}}(y_{t-1}) - g_{x_{t-2}}^{\text{opt}}) + c_2\|\nabla F(x_{t-2})\|^2\right] \\
&\leq \frac{1}{\sigma}\left[(c_1 + c_2c_3)cr^{t-1} + c_2cr^{t-1}\right] \\
&\leq \frac{1}{\sigma}c(c_1 + c_2 + c_2c_3)r^{t-1} \\
&\leq cr^t.
\end{aligned}$$

For the second invariant (22), we have already handled the case of $t = 2$, and so assume that $t \geq 3$. By Lemmas C.1 and C.2, and by the choice of σ, we have that

$$\begin{aligned}
\|\nabla F(x_{t-1})\| &\leq (1+\gamma)\|\nabla f_{x_{t-2}}(x_{t-1})\| + \gamma\|\nabla F(x_{t-2})\| \\
&\leq (1+\gamma)\sqrt{c_3(g_{x_{t-2}}(y_{t-1}) - g_{x_{t-2}}^{\text{opt}})} + \gamma\|\nabla F(x_{t-2})\| \\
&\leq (1+\gamma)\sqrt{c_3\frac{1}{\sigma}(g_{x_{t-2}}(y_{t-2}) - g_{x_{t-2}}^{\text{opt}})} + \gamma\|\nabla F(x_{t-2})\| \\
&\leq (1+\gamma)\sqrt{c_3\frac{1}{\sigma}\left[(c_1 + c_2c_3)(g_{x_{t-3}}(y_{t-2}) - g_{x_{t-3}}^{\text{opt}}) + c_2\|\nabla F(x_{t-3})\|^2\right]} + \gamma\|\nabla F(x_{t-2})\| \\
&\leq (1+\gamma)\sqrt{c_3\frac{1}{\sigma}(c_1 + c_2 + c_2c_3)cr^{t-2}} + \gamma\sqrt{cr^{t-1}} \\
&\leq (\tau-\gamma)\sqrt{cr^{t-1}} + \gamma\sqrt{cr^{t-1}} \\
&\leq \tau\sqrt{cr^{t-1}} \\
&\leq \sqrt{cr^t},
\end{aligned}$$

where the final inequality is due to the choice of $r \geq \tau^2$. This completes the inductive argument. □

Equipped with Proposition C.4, we translate it into a statement about the convergence rate of Dual APPA in error, rather than in gradient norm, to prove Theorem 2.10:

Proof of Theorem 2.10. Define γ, τ, r, and σ as in Proposition C.4. In the execution of Dual APPA (Algorithm 3), at each iteration $t \geq 1$, we claim that

$$F(x_t) - F^{\text{opt}} \leq \frac{2nR^2L}{\mu}\epsilon_0 r^{t+1}, \tag{24}$$

where $\epsilon_0 = F(x_0) - F^{\text{opt}}$. Consequently, for any $\epsilon > 0$, in order to achieve ϵ error in Dual APPA, it suffices to take a number of iterations

$$T \geq \frac{1}{1-r}\left(\log\left(\frac{2nR^2L}{\mu}\right) + \log\frac{\epsilon_0}{\epsilon}\right).$$

The first part of the claim follows directly from Proposition C.4, using the smoothness and strong convexity of F. The second part follows from a calculation sufficient to make the right hand side of (24) smaller than a given ϵ. □

D. Convergence analysis of Accelerated APPA

The goal of this section is to establish the convergence rate of Algorithm 2, Accelerated APPA, and prove Theorem 2.6. Note that the results in this section use nothing about the structure of F other than strong convexity and thus they apply to the general ERM problem (2); they are written in greater generality to make this distinction clear.

Aspects of the proofs in this section bear resemblance to the analysis in Shalev-Shwartz & Zhang (2014), which achieves similar results in a more specialized setting.

The rest of this section is structured as follows:

- In Lemma D.1 we show that applying a primal oracle to the inner minimization problem gives us a quadratic lower bound on $F(x)$.
- In Lemma D.2 we use this lower bound to construct a series of lower bounds for the main objective function f, and accelerate the APPA algorithm, comprising the bulk of the analysis.
- In Lemma D.3 we show that the requirements of Lemma D.2 can be met by using a primal oracle that decreases the error by a constant factor.
- In Lemma D.4 we analyze the initial error requirements of Lemma D.2.
- In Lemma D.5 we provide an auxiliary lemma that combines two quadratic functions that we use in the earlier proofs.

The proof of Theorem 2.6 follows immediately from these lemmas.

Lemma D.1. *Let $f \in \mathbb{R}^n \to \mathbb{R}$ be μ-strongly convex and let $f_{x_0,\lambda}(x) \stackrel{\text{def}}{=} f(x) + \frac{\lambda}{2}\|x - x_0\|_2^2$. Suppose x^+ satisfies.*

$$f_{x_0,\lambda}(x^+) \leq \min_{x\in\mathbb{R}^n} f_{x_0,\lambda}(x) + \epsilon,$$

Then for $\mu' \stackrel{\text{def}}{=} \mu/2$, $g \stackrel{\text{def}}{=} \lambda(x_0 - x^+)$, and all $x \in \mathbb{R}^n$ we have

$$f(x) \geq f(x^+) - \frac{1}{2\mu'}\|g\|^2 + \frac{\mu'}{2}\left\|x - \left(x_0 - \left(\frac{1}{\mu'} + \frac{1}{\lambda}\right)g\right)\right\|_2^2 - \frac{\lambda + 2\mu'}{\mu'}\epsilon.$$

Note that as $\mu' = \mu/2$ we are only losing a factor of 2 in the strong convexity parameter for our lower bound. This allows us to account for the error without loss in our ultimate convergence rates.

Proof. Let $x^{\text{opt}} = \text{argmin}_{x\in\mathbb{R}^n} f_{x_0,\lambda}(x)$. Since f is μ-strongly convex clearly $f_{x_0,\lambda}$ is $\mu + \lambda$ strongly convex and by Lemma B.1

$$f_{x_0,\lambda}(x) - f_{x_0,\lambda}(x^{\text{opt}}) \geq \frac{\mu + \lambda}{2}\|x - x^{\text{opt}}\|_2^2. \tag{25}$$

By Cauchy-Schwartz we know

$$\frac{\lambda + \mu'}{2}\|x - x^+\|_2^2 \leq \frac{\lambda + \mu'}{2}\left(\|x - x^{\text{opt}}\|_2^2 + \|x^{\text{opt}} - x^+\|_2^2\right) + \frac{\mu'}{2}\|x - x^{\text{opt}}\|_2^2 + \frac{(\lambda + \mu')^2}{2\mu'}\|x^{\text{opt}} - x^+\|_2^2,$$

which implies

$$\frac{\mu+\lambda}{2}\|x-x^{\text{opt}}\|_2^2 \geq \frac{\lambda+\mu'}{2}\|x-x^+\|_2^2 - \frac{\lambda+\mu'}{\mu'}\cdot\frac{\lambda+\mu}{2}\|x^{\text{opt}}-x^+\|_2^2.$$

On the other hand, since $f_{x_0,\lambda}(x^+) \leq f_{x_0,\lambda}(x^{\text{opt}}) + \epsilon$ by (25) we have $\frac{\lambda+\mu}{2}\|x^+ - x^{\text{opt}}\|_2^2 \leq \epsilon$ and therefore

$$\begin{aligned}
f_{x_0,\lambda}(x) - f_{x_0,\lambda}(x^+) &\geq f_{x_0,\lambda}(x) - f_{x_0,\lambda}(x^{\text{opt}}) - \epsilon \\
&\geq \frac{\mu+\lambda}{2}\|x-x^{\text{opt}}\|_2^2 - \epsilon \\
&\geq \frac{\lambda+\mu'}{2}\|x-x^+\|_2^2 - \frac{\lambda+\mu'}{\mu'}\cdot\frac{\lambda+\mu}{2}\|x^{\text{opt}}-x^+\|_2^2 - \epsilon \\
&\geq \frac{\lambda+\mu'}{2}\|x-x^+\|_2^2 - \frac{\lambda+2\mu'}{\mu'}\epsilon.
\end{aligned}$$

Now since

$$\|x-x^+\|_2^2 = \|x-x_0+\frac{1}{\lambda}g\|_2^2 = \|x-x_0\|^2 + \frac{2}{\lambda}\langle g, x-x_0\rangle + \frac{1}{\lambda^2}\|g\|_2^2,$$

and using the fact that $f_{x_0,\lambda}(x) = f(x) + \frac{\lambda}{2}\|x-x_0\|_2^2$, we have

$$f(x) \geq f(x^+) + \left[\frac{1}{\lambda} + \frac{\mu'}{2\lambda^2}\right]\|g\|_2^2 + \left(1+\frac{\mu'}{\lambda}\right)\langle g, x-x_0\rangle + \frac{\mu'}{2}\|x-x_0\|^2 - \frac{\lambda+2\mu'}{\mu'}\epsilon.$$

The right hand side of the above equation is a quadratic function. Looking at its gradient with respect to x we see that it obtains its minimum when $x = x_0 - \left(\frac{1}{\mu'}+\frac{1}{\lambda}\right)g$ and has a minimum value of $f(x^+) - \frac{1}{2\mu'}\|g\|_2^2 - \frac{\lambda+2\mu'}{\mu'}\epsilon$. □

Lemma D.2. *Let $f:\mathbb{R}^n\to\mathbb{R}$ be μ-strongly convex and suppose that in each iteration t we have $\psi_t \stackrel{\text{def}}{=} \psi_t^{opt} + \frac{\mu'}{2}\|x-v^{(t)}\|_2^2$ such that $f(x)\geq\psi_t(x)$ for all x. Then if we let $\rho \stackrel{\text{def}}{=} \frac{\mu'+\lambda}{\mu'}$, $f_{y,\lambda}(x) \stackrel{\text{def}}{=} f(x) + \frac{\lambda}{2}\|y-x\|_2^2$ for $\lambda \geq 3\mu'$ and*

- $y^{(t)} \stackrel{\text{def}}{=} \frac{1}{1+\rho^{-1/2}}x^{(t)} + \frac{\rho^{-1/2}}{1+\rho^{-1/2}}v^{(t)}$,
- $\mathbb{E}[f_{y^{(t)},\lambda}(x^{(t+1)})] - f_{y^{(t)},\lambda}^{opt} \leq \frac{\rho^{-3/2}}{4}(f(x^{(t)}) - \psi_t^{opt})$,
- $g^{(t)} \stackrel{\text{def}}{=} \lambda(y^{(t)} - x^{(t+1)})$,
- $v^{(t+1)} \stackrel{\text{def}}{=} (1-\rho^{-1/2})v^{(t)} + \rho^{-1/2}\left[y^{(t)} - \left(\frac{1}{\mu'}+\frac{1}{\lambda}\right)g^{(t)}\right]$.

We have that

$$\mathbb{E}[f(x^{(t)}) - \psi_t^{opt}] \leq \left(1 - \frac{\rho^{-1/2}}{2}\right)^t (f(x_0) - \psi_0^{opt}).$$

Proof. Regardless of how $y^{(t)}$ is chosen we know by Lemma D.1 that for $\gamma = 1 + \frac{\mu'}{\lambda}$ and all $x\in\mathbb{R}^n$

$$f(x) \geq f(x^{(t+1)}) - \frac{1}{2\mu'}\|g^{(t)}\|_2^2 + \frac{\mu'}{2}\left\|x - \left(y^{(t)} - \frac{\gamma}{\mu'}g^{(t)}\right)\right\|_2^2 - \frac{\lambda+2\mu'}{\mu'}\left(f_{y^{(t)},\lambda}(x^{(t+1)}) - f_{y^{(t)},\lambda}^{\text{opt}}\right). \tag{26}$$

Thus, for $\beta = 1 - \rho^{-1/2}$ we can let

$$\begin{aligned}
\psi_{t+1}(x) &\stackrel{\text{def}}{=} \beta\psi_t(x) + (1-\beta)\left[f(x^{(t+1)}) - \frac{1}{2\mu'}\|g^{(t)}\|_2^2 + \frac{\mu'}{2}\|x - \left(y^{(t)} - \frac{\gamma}{\mu'}g^{(t)}\right)\|_2^2\right. \\
&\quad \left. - \frac{\lambda + 2\mu'}{\mu'}(f_{y^{(t)},\lambda}(x^{(t+1)}) - f^{\text{opt}}_{y^{(t)},\lambda})\right] \\
&= \beta\left[\psi_t^{\text{opt}} + \frac{\mu'}{2}\|x - v^{(t)}\|_2^2\right] + (1-\beta)\left[f(x^{(t+1)}) - \frac{1}{2\mu'}\|g^{(t)}\|_2^2 + \frac{\mu'}{2}\|x - \left(y^{(t)} - \frac{\gamma}{\mu'}g^{(t)}\right)\|_2^2\right. \\
&\quad \left. - \frac{\lambda + 2\mu'}{\mu'}(f_{y^{(t)},\lambda}(x^{(t+1)}) - f^{\text{opt}}_{y^{(t)},\lambda})\right] \\
&= \psi_{t+1}^{\text{opt}} + \frac{\mu'}{2}\|x - v^{(t+1)}\|_2^2.
\end{aligned}$$

where in the last line we used Lemma D.5. Again, by Lemma D.5 we know that

$$\begin{aligned}
\psi_{t+1}^{\text{opt}} &= \beta\psi_t + (1-\beta)\left(f(x^{(t+1)}) - \frac{1}{2\mu'}\|g^{(t)}\|_2^2 - \frac{\lambda + 2\mu'}{\mu'}(f_{y^{(t)},\lambda}(x^{(t+1)}) - f^{\text{opt}}_{y^{(t)},\lambda})\right) \\
&\quad + \beta(1-\beta)\frac{\mu'}{2}\|v^{(t)} - \left(y^{(t)} - \frac{\gamma}{\mu'}g^{(t)}\right)\|_2^2 \\
&\geq \beta\psi_t + (1-\beta)f(x^{(t+1)}) - \frac{(1-\beta)^2}{2\mu'}\|g^{(t)}\|_2^2 + \beta(1-\beta)\gamma\left\langle g^{(t)}, v^{(t)} - y^{(t)}\right\rangle \\
&\quad - \frac{(1-\beta)(\lambda + 2\mu')}{\mu'}(f_{y^{(t)},\lambda}(x^{(t+1)}) - f^{\text{opt}}_{y^{(t)},\lambda}).
\end{aligned}$$

In the second step we used the following fact:

$$-\frac{1-\beta}{2\mu'} + \beta(1-\beta)\frac{\mu'}{2}\cdot\frac{\gamma^2}{\mu'} = \frac{1-\beta}{2\mu'}(-1 + \beta\gamma^2) \geq -\frac{(1-\beta)^2}{2\mu'}.$$

Furthermore, expanding the term $\frac{\mu}{2}\|(x - y^{(t)}) + \frac{\gamma}{\mu}g^{(t)}\|_2^2$ and instantiating x with $x^{(t)}$ in (26) yields

$$f(x^{(t+1)}) \leq f(x^{(t)}) - \frac{1}{\lambda}\|g^{(t)}\|_2^2 + \gamma\left\langle g^{(t)}, y^{(t)} - x^{(t)}\right\rangle + \frac{\lambda + 2\mu'}{\mu'}(f_{y^{(t)},\lambda}(x^{(t+1)}) - f^{\text{opt}}_{y^{(t)},\lambda}).$$

Consequently we know

$$\begin{aligned}
f(x^{(t+1)}) - \psi_{t+1}^{\text{opt}} &\leq \beta[f(x^{(t)}) - \psi_t^{\text{opt}}] + \left[\frac{(1-\beta)^2}{2\mu'} - \frac{\beta}{\lambda}\right]\|g^{(t)}\|_2^2 + \gamma\beta\left\langle g^{(t)}, y^{(t)} - x^{(t)} - (1-\beta)(v^{(t)} - y^{(t)})\right\rangle \\
&\quad + \frac{(\lambda + 2\mu')}{\mu'}(f_{y^{(t)},\lambda}(x^{(t+1)}) - f^{\text{opt}}_{y^{(t)},\lambda})
\end{aligned}$$

Note that we have chosen $y^{(t)}$ so that the inner product term equals 0, and we choose $\beta = 1 - \rho^{-1/2} \geq \frac{1}{2}$ which ensures

$$\frac{(1-\beta)^2}{2\mu'} - \frac{\beta}{\lambda} \leq \frac{1}{2(\mu' + \lambda)} - \frac{1}{2\lambda} \leq 0.$$

Also, by assumption we know $\mathbb{E}[f_{y^{(t)},\lambda}(x^{(t+1)}) - f^{\text{opt}}_{y^{(t)},\lambda}] \leq \frac{\rho^{-3/2}}{4}(f(x^{(t)}) - \psi_t^{\text{opt}})$, which implies

$$\mathbb{E}[f(x^{(t+1)}) - \psi_{t+1}^{\text{opt}}] \leq \left(\beta + \frac{(\lambda + 2\mu')}{\mu'}\cdot\frac{\rho^{-3/2}}{4}\right)(f(x^{(t)}) - \psi_t^{\text{opt}}) \leq (1 - \rho^{-1/2}/2)(f(x^{(t)}) - \psi_t^{\text{opt}}).$$

In the final step we are using the fact that $\frac{\lambda + 2\mu'}{\mu'} \leq 2\rho$ and $\rho \geq 1$. □

Lemma D.3. *Under the setting of Lemma D.2, we have* $f_{y^{(t)},\lambda}(x^{(t)}) - f^{opt}_{y^{(t)},\lambda} \leq f(x^{(t)}) - \psi_t^{opt}$. *In particular in order to achieve* $\mathbb{E}[f_{y^{(t)},\lambda}(x^{(t+1)}) \leq \frac{\rho^{-3/2}}{8}(f(x^{(t)}) - \psi_t^{opt})$ *we only need an oracle that shrinks the function error by a factor of* $\frac{\rho^{-3/2}}{8}$ *(in expectation).*

Proof. We know

$$f_{y^{(t)},\lambda}(x^{(t)}) - f(x^{(t)}) = \frac{\lambda}{2}\|x^{(t)} - y^{(t)}\|_2^2 = \frac{\lambda}{2} \cdot \frac{\rho^{-1}}{(1+\rho^{-1/2})^2}\|x^{(t)} - v^{(t)}\|_2^2.$$

We will try to show the lower bound $f^{\text{opt}}_{y^{(t)},\lambda}$ is larger than ψ_t^{opt} by the same amount. This is because for all x we have

$$f_{y^{(t)},\lambda}(x) = f(x) + \frac{\lambda}{2}\|x - y^{(t)}\|_2^2 \geq \psi_t^{\text{opt}} + \frac{\mu'}{2}\|x - v^{(t)}\|_2^2 + \frac{\lambda}{2}\|x - y^{(t)}\|_2^2.$$

The RHS is a quadratic function, whose optimal point is at $x = \frac{\mu' v^{(t)} + \lambda y^{(t)}}{\mu' + \lambda}$ and whose optimal value is equal to

$$\psi_t^{\text{opt}} + \frac{\lambda}{2}\left(\frac{\mu'}{\mu'+\lambda}\right)^2\|v^{(t)} - y^{(t)}\|_2^2 + \frac{\mu'}{2}\left(\frac{\lambda}{\mu+\lambda}\right)^2\|v^{(t)} - y^{(t)}\|_2^2 = \psi_t^{\text{opt}} + \frac{\mu'\lambda}{2(\mu'+\lambda)} \cdot \frac{1}{(1+\rho^{-1/2})^2}\|x^{(t)} - v^{(t)}\|_2^2.$$

By definition of ρ^{-1}, we know $\frac{\mu'\lambda}{2(\mu'+\lambda)} \cdot \frac{1}{(1+\rho^{-1/2})^2}\|x^{(t)} - v^{(t)}\|_2^2$ is exactly equal to $\frac{\lambda}{2} \cdot \frac{\rho^{-1}}{(1+\rho^{-1/2})^2}\|x^{(t)} - v^{(t)}\|_2^2$, therefore $f_{y^{(t)},\lambda}(x^{(t)}) - f^{\text{opt}}_{y^{(t)},\lambda} \leq f(x^{(t)}) - \psi_t^{\text{opt}}$. □

Remark In the next lemma we show that moving to the regularized problem has the same effect on the primal function value and the lower bound. This is a result of the choice of β in the proof of Lemma D.2. However this does not mean the choice of β is very fragile, we can choose any β' that is between the current β and 1; the influence to this lemma will be that the increase in primal function becomes smaller than the increase in the lower bound (so the lemma continues to hold).

Lemma D.4. *Let* $\psi_0^{opt} = f(x^{(0)}) - \frac{\lambda+2\mu'}{\mu'}(f(x^{(0)}) - f^{opt})$, *and* $v^{(0)} = x^{(0)}$, *then* $\psi_0 \overset{\text{def}}{=} \psi_0^{opt} + \frac{\mu'}{2}\|x - v_0\|^2$ *is a valid lower bound for* f. *In particular when* $\lambda = LR^2$ *then* $f(x^{(0)}) - \psi_0^{opt} \leq 2\kappa(f(x^{(0)}) - f^{opt})$.

Proof. This lemma is a direct corollary of Lemma D.1 with $x^+ = x^{(0)}$. □

Lemma D.5. *Suppose that for all* x *we have*

$$f_1(x) \overset{\text{def}}{=} \psi_1 + \frac{\mu}{2}\|x - v_1\|_2^2 \text{ and } f_2(x) = \psi_2 + \frac{\mu}{2}\|x - v_2\|_2^2$$

then

$$\alpha f_1(x) + (1-\alpha)f_2(x) = \psi_\alpha + \frac{\mu}{2}\|x - v_\alpha\|_2^2$$

where

$$v_\alpha = \alpha v_1 + (1-\alpha)v_2 \quad \text{and} \quad \psi_\alpha = \alpha\psi_1 + (1-\alpha)\psi_2 + \frac{\mu}{2}\alpha(1-\alpha)\|v_1 - v_2\|_2^2$$

Proof. Setting the gradient of $\alpha f_1(x) + (1-\alpha)f_2(x)$ to 0 we know that v_α must satisfy

$$\alpha\mu\,(v_\alpha - v_1) + (1-\alpha)\mu\,(v_\alpha - v_2) = 0$$

and thus

$$v_\alpha = \alpha v_1 + (1-\alpha)v_2.$$

Finally,

$$\begin{aligned}
\psi_\alpha &= \alpha\left[\psi_1 + \frac{\mu}{2}\|v_\alpha - v_1\|_2^2\right] + (1-\alpha)\left[\psi_2 + \frac{\mu}{2}\|v_\alpha - v_2\|_2^2\right] \\
&= \alpha\psi_1 + (1-\alpha)\psi_2 + \frac{\mu}{2}\left[\alpha(1-\alpha)^2\|v_2 - v_1\|_2^2 + (1-\alpha)\alpha^2\|v_2 - v_1\|_2^2\right] \\
&= \alpha\psi_1 + (1-\alpha)\psi_2 + \frac{\mu}{2}\alpha(1-\alpha)\|v_1 - v_2\|_2^2
\end{aligned}$$

□