# 13 Dual methods

Last lecture we saw that to any convex optimization problem with an affine constraint, one can associate a Lagrange dual. The optimal value of the dual problem is equal to the optimal value of the original, primal, problem under some mild conditions (e.g., Slater's condition). In some cases, the dual problem has a structure that is more amenable to algorithms than the original, primal, problem. We explore the possibility of applying various optimization methods to the *dual* problem.

Consider an optimization problem of the form

$$\min_{x\in\mathbb{R}^n} \quad f(x)+h(Ax). \tag{1}$$

Problems of the type (1) arise often in *statistics* and *inverse problems* where, e.g., $f(x)$ is a data-fidelity term, and $h(Ax)$ is a regularization term. Typically $f$ is smooth and strongly convex, and $h$ is nonsmooth with a simple prox. Note that even if $\mathbf{prox}_h$ is easy to compute, computing $\mathbf{prox}_{h\circ A}$ can be hard.

**Example** (Signal denoising using total variation)**.** *Consider the problem of denoising a 1D signal $u\in\mathbb{R}^n$ with total-variation regularization*

$$\min_{x\in\mathbb{R}^n} \quad \sum_{i=1}^n (x_i-u_i)^2+\lambda\sum_{i=1}^{n-1}|x_{i+1}-x_i|.$$

*This problem can be put in the form (1) with $f(x)=\|x-u\|_2^2$, $h(x)=\|x\|_1$ and $A$ is the discrete difference operator.*

We can rewrite problem (1) as

$$\min_{x,y} \quad f(x)+h(y) \quad \text{subject to} \quad y=Ax.$$

The Lagrangian is

$$L(x,y,z)=f(x)+h(y)+\langle z,Ax-y\rangle \tag{2}$$

and the dual function is

$$\begin{aligned} g(z)=\min_{x,y} L(x,y,z) &= \min_{x,y}\{f(x)+\langle z,Ax\rangle+h(y)-\langle z,y\rangle\} \\ &= \min_x\{f(x)+\langle z,Ax\rangle\}+\min_y\{h(y)-\langle z,y\rangle\} \\ &= -f^*(-A^Tz)-h^*(z). \end{aligned} \tag{3}$$

So the dual problem is

$$\max_{z\in\mathbb{R}^n} -f^*(-A^Tz)-h^*(z). \tag{4}$$

**Proximal gradient to dual** If $f$ is strongly convex then $z \mapsto f^*(-A^T z)$ is smooth and its gradient has Lipschitz constant $\|A\|^2/m$ where $m$ is the strong convexity parameter of $f$. One can apply the proximal gradient method to the dual problem (4). This gives the iteration rule:

$$z_{k+1} = \mathbf{prox}_{th^*}(z_k + tA\nabla f^*(-A^T z_k)). \tag{5}$$

where $t > 0$ is the time step. We can simplify the iteration rule using the definitions of $\nabla f^*$ and of **prox**. Indeed, we saw before that since $f$ is strongly convex

$$\nabla f^*(y) = \operatorname*{argmax}_x \{\langle y, x\rangle - f(x)\} = \operatorname*{argmin}_x \{f(x) - \langle y, x\rangle\}.$$

Thus Equation (5) takes the form

$$\begin{aligned} \hat{x} &= \operatorname*{argmin}_x \{f(x) + \langle z_k, Ax\rangle\} \\ z_{k+1} &= \mathbf{prox}_{th^*}(z_k + tA\hat{x}). \end{aligned}$$

We can further simplify the equations above using Moreau's identity (see exercise sheet 3) which tells us that $\mathbf{prox}_{\phi^*}(x) = x - \mathbf{prox}_{\phi}(x)$ for any closed convex function $\phi$. With $\phi = th^*$ we get $\phi^*(y) = (th^*)^*(y) = th(y/t)$ (check!). Also one can verify that $\mathbf{prox}_{th(\cdot/t)}(x) = t\,\mathbf{prox}_{t^{-1}h}(x/t)$. At the end, after all simplifications, the proximal gradient method applied to the dual problem (4) takes the form:

$$\text{Proximal gradient applied to dual pb (4):} \quad \begin{cases} \hat{x} = \operatorname*{argmin}_x \{f(x) + \langle z_k, Ax\rangle\} \\ \hat{y} = \operatorname*{argmin}_y \left\{h(y) - \langle z_k, y\rangle + \dfrac{t}{2}\|A\hat{x} - y\|_2^2\right\} \\ z_{k+1} = z_k + t(A\hat{x} - \hat{y}). \end{cases} \tag{6}$$

In the signal denoising example (where $f(x) = \|x - u\|_2^2$ and $h(z) = \|z\|_1$) note that $\hat{x}$ and $\hat{y}$ can be computed easily with a closed-form expression.

*Comparison with dual ascent:* It is instructive to compare (6) to a *subgradient ascent* method applied to the dual problem (4). Using the expression of the dual function in (3) dual ascent takes the form

$$\text{Subgradient ascent applied to dual pb (4):} \quad \begin{cases} \hat{x} = \operatorname*{argmin}_x \{f(x) + \langle z_k, Ax\rangle\} \\ \hat{y} = \operatorname*{argmin}_y \{h(y) - \langle z_k, y\rangle\} \\ z_{k+1} = z_k + t_k(A\hat{x} - \hat{y}). \end{cases} \tag{7}$$

Unless $f$ and $h$ are both strongly convex, the dual function $g(z)$ in (3) is not going to be smooth; this means that the step sizes $t_k$ has to be decreasing, and in general the above subgradient ascent is going to be very slow.

**Augmented Lagrangian method** It is also instructive to compare (6) with the augmented Lagrangian method, which does not require any strong convexity assumption on $f$ or $h$: observe that the original problem can be written as

$$\min_{x,y} \left\{ f(x) + h(y) + \frac{t}{2}\|Ax - y\|_2^2 \quad : \quad Ax = y \right\}$$

where $t > 0$. The Lagrangian of this problem is

$$L_t(x, y, z) = f(x) + h(y) + \langle z, Ax - y\rangle + \frac{t}{2}\|Ax - y\|_2^2.$$

This is known as the *augmented Lagrangian* of the original problem. The dual function is

$$g_t(z) = \min_{x,y} \left\{ f(x) + h(y) + \langle z, Ax - y\rangle + \frac{t}{2}\|Ax - y\|_2^2 \right\}. \tag{8}$$

Because of the quadratic term $\frac{t}{2}\|Ax-y\|_2^2$, one can show that $g$ is $(1/t)$-smooth[1], and that $\nabla g_t(z) = A\hat{x} - \hat{y}$ where $(\hat{x}, \hat{y})$ are minimizers in (8). The augmented Lagrangian method corresponds to a gradient ascent on $g_t$, i.e., it takes the form

$$\text{Augmented Lagrangian method:} \quad \begin{cases} (\hat{x}, \hat{y}) = \operatorname*{argmin}_{x,y} \left\{ f(x) + h(y) + \langle z_k, Ax - y\rangle + \frac{t}{2}\|Ax - y\|_2^2 \right\} \\ z_{k+1} = z_k + t(A\hat{x} - \hat{y}). \end{cases} \tag{9}$$

[1]Indeed, note that by introducing $v = Ax - y$ we can write $g_t(z) = \min_v \{\min_x\{f(x) + h(Ax - v)\} + (t/2)\|v\|_2^2 + z^T v\} = -\psi^*(-z)$ where $\psi(v) = (t/2)\|v\|_2^2 + \min_x(f(x) + h(Ax - v))$ is $t$-strongly convex.